Notes from OHDSI OMOP Hadoop Working Group

10/28/16

[sdolley@cloudera.com](mailto:sdolley@cloudera.com)

RAW NOTES AS TAKEN – no changes.

Roll Call

* Albert “Al” Pivonka
* Aaron Galaznik
* Brack Coalson
* David Ficenec (pronouned Fis-nek)
* Lee Evans
* Malcolm McRoberts
* Nagakalyana Eskala
* Steve Lyman
* Tom White
* Vojtech Huser
* Charity Hilton
* Others (not on webex, but on phone): \_\_\_\_

Not in attendance but interested: TBD, Frank de Falco

From Patrick: “I don't think everyone is aligned on what they'd use Hadoop for, or why.  Some are looking for etl platform, some are looking for storage/warehouse, some are looking for analytics engine.  All of those use cases are viable, but they are quite distinct.  One of the first orders of business for the workgroup should be to align on language and use cases”

Question: is Hadoop only for OMOP CDM? Or are there other OHDSI projects that would leverage Hadoop

Use Cases one might use Hadoop for? (some of these are use cases and some of these are benefits of Hadoop)

* Doing things outside of SQL (might be ‘support for code based data access rather than SQL based)
* ETL platform (this was re-iterated).  How does this relate to cloud vs on-prem.
* Storage/warehouse
* Analytics engine
* Delivering speed and efficiency in analyzing claims and other data
* Different data sources can be compared via code to see how they are different
* Speed might be based
* From existing member in using Hadoop for NLP (non OMOP)
* SQL Render (does Hadoop support SQL-92?) Running OHDSI tools.  Someone tried to port to Google BigQuery and ran into problems with temp tables.  CSV file inside \_\_ and drivers. If you are good at R and have Hadoop, can make change to csv file and it will work?
* OMOP Common Data Model, use of that solely as a data model, untethered to any metrics that OHDSI/OMOP has today
* Use cases for graph databases? Interesting in the concept and terminology space.
* Use cases for search?  Has Google released their search engine?  SOLR is part of Hadoop ecosystem (based on Lucene which everyone uses). Regenstrief puts structured in SOLR and runs it without having Hadoop stack.
* Sensor data - model would need to accommodate it first perhaps

Notes and questions:

* OHDSI stack centered on SQL.  Can’t do it all in R.
* SQL Render engine uses R (and wrapping Java functions). Uses csv to translate SQL Server dialect to other dialects.
* Typical test is run full Achilles library, if Achilles will run then it is done
* One could add two different new translations (?) in csv file, one for Hive and one for Impala?
* Could one, on a query by query basis, choose between Impala and Hive.  But data stored once in HDFS.  Table structures would be stored once in Hive metastorm.
* Look at other open source projects that serve same function: HDB (hortonworks), Hbase, Tez, Splice Machines (not open source?),  - whichever ones are worked on probably need to support SparkSQL.
* Options for not using HDFS, access Hive Metastore for table definitions and use R
* For ad hoc, use Impala and SparkSQL, and for batch use Hive.
* YARN and MESOS are different.  We may not need to ‘handle’ this.
* Is there a need to support Hive queries (beyond just using Hive metastore)?
* Taking queries and analysis that we have, some rewriting? Cannot natively take R code today and make it SparkR? Support of data frames might be intermediate goal?
* Note- existing R code is not pure R, it has SQL within it.  We can’t escape it, so need to embrace it.
* What is roadmap for Cloudera in Hive and Impala to improve SQL? From Tom: general direction that Hive and Impala are adding more functionality (e.g. windowing functions) over time. Cloudera has been assuming HDFS would be storage, new one coming called Kudu, allows updating a database like normal db’s.  Hadoop db’s today cannot update existing rows, Kudu does and available immediately.
* Vojtech: first: making existing tools work on Hadoop; second: Hadoop is opening up new tools
* What is roadmap for SparkR in CDH? Also for SparklyR what is the roadmap vis a vis CDH?
* Use of notebooks to re-create some of these tools? [some words I didn’t catch]
* Need to support S3.  If ETL is hitting S3 as the source.  Appending new data as it comes in.  Dataframes in Apache Spark (using for ETL) has worked well in some ways.
* We need to tell people what file formats work best and worst per data/etc.  “Cuts a lot of wasted”.
* Looking at non-Hadoop ‘tools’ to improve speed, performance. Patient level prediction or Cohort, optimization.  To think about: if our goal is performance, than widen the focus to include other areas outside of Hadoop
* Lee: creating pre-aggregated data sets (such as frequently used cohorts or cohorts of interest), there is a need there and could be a benefit. Could tie to Achilles.
* Think about data that is not traditionally healthcare data, what is roadmap to merge different kinds of data (e.g. linked data?).  GA Tech is thinking about sensor data e.g.
* How can we get access for members to a cluster? Quickstart VM allows you to run a cluster on a machine.

What is done by any member already

* Tom White did something to create Hive tables that map to OMOP CDM.  Hive has a metastore, so table definitions live there.  Impala hits Hive metastore. SQL hits Impala, not Hive.
* Patrick Ryan committed that 10/27/16
* Looked at windowing functions in Impala (this troubled Google Big Query I think?)
* Tom had to convert date/time into strings.  Impala has timestamp but OMOP not at timestamp all the time.  Maybe this would make it slower? Maybe there is a way to solve

NEXT STEPS

1. Post these notes
2. Reformat these notes
3. Get a list of things to vote on
4. Get votes
5. Get list of coding
6. Next time record meeting!!